#381 Compression Coding Model for Password Guessability

Email notification

Select to receive email on updates to reviews and comments.

▼ PC conflicts

Siqi Ma

Rejected

Submission (3.8MB) Dec 7, 2023, 11:45:11 PM AoE · d4381fba

▼ **Abstract**

Passwords are currently the most commonly used method of authentication, but there is a constant struggle between password security and guessability. This is a concern in both the cybersecurity and information theory communities. Unfortunately, these two fields of password research tend to operate separately with little cross-pollination. There is barely help of information theory to study and solve the real-world password issues in current research.

In this work, we, *for the first time*, introduce a compression coding model for password research, and demonstrate the inherent relationship between data compression and password guessability. Specifically, we explain password guessing behavior using data compression theory and measure the guessability of PIN codes (a special kind of password) and regular website passwords by coding methods. The results obtained are compared to conventional password guessability metrics. Furthermore, we apply data compression theory to provide a concise proof and concrete understanding of a classical guesswork-entropy relation in password research. This relation was first proposed by Massey in a complex manner of convex optimization, but our approach simplifies it and demonstrates the consistency between the two fields.

The consistency helps us better understand the security of real-world passwords. Our

▶ **Authors** (anonymous)

X. Lu, D. Wang, Z. Hou [details]

quantitative results demonstrate the impact of an attacker's capability on password guessability. For instance, when attackers have access to the Dodonew dataset, the guessability of 4-digit PINs in the CSDN dataset is -10.5 bits, whereas it decreases to -11.3 bits if only the Yahoo dataset information is available.

Furthermore, we employ our compression coding model to compress the guessing process. This involves cracking more passwords with fewer guesses by utilizing various guessing methods. Our guessing-compression algorithm (CompGuess) combines multiple guessing methods, and we conduct experiments corresponding to 13 large-scale password datasets to validate its feasibility. In online guessing, CompGuess outperforms the best-performing guessing method by 17.4% within 10^4 guesses (avg. 12.5% for targeted password guessing). Besides, CompGuess achieves an average compression ratio (offline guessing cost-related metric, Sec. 5.4) of 53.8% for offline guessing, showcasing its superiority in compressing the guessing process. Our work underscores the practicality and utility of our compression coding model in password research.

	OveMer	RevExp
Review #381A	3	3
Review #381B	3	2
Review #381C	3	1
Review #381D	2	1
Review #381E	3	3

You are an **author** of this submission.

Edit submission

Reviews in plain text

Review #381A

Overall merit

3. Weak accept

Reviewer expertise

3. Knowledgeable

Paper summary

This paper reveals the connection between the password guessing and data compression, providing fresh insights into the security of human passwords in real-world situations. The authors noticed that password guessing has been studied in the fields of cybersecurity and information theory but with little cross-pollination. They thus formulate the password guessing process using a compression coding model. The models reframes the task of password guessing as the construction of a strategy for determining the order of guesses. Their proposed CompGuess approach applied with adaptive arithmetic coding effectively tackles the challenge of synthesizing multiple guessing methods within the compression process of guessing, revealing the security of passwords in real-world scenarios.

Comments for authors

Strengths:

+ Valuable insights were offered by formulating the password guessing process as a data compression problem, providing an innovative perspective to better understand the guessability of passwords.

+ The paper defines the password probability space model, attack model and relevant metrics for the compression coding model in the context of the information theory. The definitions of the three hierarchical attackers and the quantification of the two gaps can serve as basis in the problem formulation for future work in this area.

+ The unification of password guessing and coding theory is supported by concrete theoretical derivations and experimental validations.

+ The paper considers real-world scenarios of password guessing attack, including the state-of-the-art guessing methods and multiple guessing scenarios (e.g., online guessing (without targeted information), targeted online guessing, and offline guessing).

Weaknesses:

– More attention is needed on the inspiration and practical deployment of this work aimed at maintaining and enhancing password security.
– Writing needs significant improvement.

Comments:

This paper investigates the issue of password guessing using techniques from information theory. It provides valuable insights and its problem modeling and formulation have practical implications for the password research community. Below, I provide a detailed review of various aspects of this paper, along with my comments and suggestions for the authors to consider.

Evaluation: The evaluation metrics used in this work for different password guessing scenarios are reasonable, but further explanations are needed. It is noted that offline guessing involves evaluation metrics related to the hash cost, while online guessing considers the number of successful guesses under a fixed number of attempts. It is recommended to provide additional explanations to address the differences between these two scenarios.

Security insights:

1. The authors address the concern of password resistance to guessing attacks in real-world scenarios. However, the paper lacks specific recommendations for improving password security. While Appendix D mentions a method for measuring password guessability, it does not provide clear explanations regarding the regarding the implementation procedures, such as whether the method should be applied on the client-side or server-side， concerns on potential user privacy disclosures.

2. The paper seems to fall short in predicting the guess number for a password, particularly for a large guess number, similar to the Monte Carlo method. Can this work propose improvements to the Monte Carlo method in the context of multiple guessing methods?

"Monte Carlo Strength Evaluation： Fast and Reliable Password Checking", CCS '15.

Presentation: The paper's presentation needs improvement. Its overall organization is commendable, but its flow is not very smooth. The same piece of information is distributed into multiple paragraphs and even sections sometimes. It is recommended to consolidate the evaluation metrics for different password guessing scenarios into a single table for better clarity.

Review #381B

Overall merit

3. Weak accept

Reviewer expertise

2. Some familiarity

Paper summary

This paper proposes a compression coding model for password research and establishes a relationship between data compression and password guessability. The paper measures password guessability using coding methods.

Review #381C

Overall merit

3. Weak accept

Reviewer expertise

1. No familiarity

Paper summary

This paper first points out the consistency between data compression and password guessability, providing a new perspective on analysing password research. Based on this new perspective, this paper also proposes a guessing-compressing algorithm which combines multiple guessing methods and can provide better guessability. This algorithm ingeniously utilises the idea of adaptive arithmetic coding.

Comments for authors

It may be better to introduce the adaptive arithmetic coding before introducing CompGuess and give a more detailed description of the algorithm, especially in the Figures.

Review #381D

Overall merit

2. Weak reject

Reviewer expertise

1. No familiarity

Paper summary

This paper investigate the suitability of data compression theory as a theoretical tool for analysis of password guessability. Specifically, the authors propose a compression coding model for password, upon which they build a compressed guessing algorithm. Results show that such models, when built on representative datasets, can accelerate the password guessing process.

Comments for authors

Thank you for submitting your work to AsiaCCS 2024. Password security is a mature area of research, but I appreciate the importance of continuing to explore how to model threats and attacks related to password guessability. The paper is outside my area of expertise, but I found the paper clearly written and attempting to provide background on both measures of password entropy and the theoretical aspects of data compression. I also appreciated that the evaluation was carried on multiple real-world datasets.

I also found issues with the paper:

- A large number of details essentials to understanding and assessing the contributions of this paper are relegated to the appendix. While this may make it easier to meet the page limit, I don't think it is an acceptable practice and significantly detract from the value of this work
- While the exploration presented here is interesting, I found that the paper fails to convincingly clarify which problem is specifically been addressed here. "Bridging the gap" is interesting and useful but should be connected to a clearly spelled out new set of capabilities that this approach provides
- Related to the point above: given its applied flavor, the paper also lacks a clearly defined threat model. Some assumptions are weak and/or unclear, such as the fact that the attacker only attempts a small number of guesses per password. Another unsubstantiated assumption is that communities of website users define specific and stable probability spaces due to their password habits.
- Also related, I found that the paper, given that it has been submitted to a generalist security conference, should make a greater effort to provide an intuition of why this approach improves guessing times. The core idea heard - as I understand - is to rerank guesses to increase chances of early matches, but it would important to discuss how dependent the effectiveness is on the specific dataset used to derive the distribution.

  - Related: I am not sure what to make of Table 1. The authors claim that the fact that the values are sort of similar shows that code length relates to guesswork, but this seems somewhat arbitrary given that some cases show significant relative differences.
- It would have been useful to discuss the different guessing methods considered in the evaluation and how they differ

Review #381E

Overall merit

3. Weak accept

Reviewer expertise

3. Knowledgeable

Paper summary

This paper applies a compression coding model (called CompGuess) to evaluate password guessability, addressing the gap between data compression theory and the analysis of password security. Through a comprehensive application of data compression principles, the study analyzes the predictability of PIN codes and website passwords, enriching our understanding of password security. By evaluating the influence of attackers' capabilities on password guessability and introducing the CompGuess algorithm, the research not only enhances guessing efficiency by integrating diverse methods but also empirically validates the algorithm's efficacy with extensive password datasets. These experiments show meaningful improvements in both online and offline password-guessing scenarios.

Comments for authors

Strengths

1. This paper introduces a novel framework that uses data compression principles to evaluate password guessability. Although the concept of using compression in this context is not entirely novel, the application presented in this study is fundamental and worthy of consideration.

2. The proposed password-guessing evaluation method offers a rigorous analysis from an information theory perspective, significantly advancing our comprehension of how data compression correlates with password guessability. The experimental results effectively demonstrate the superiority of the proposed approach over existing evaluation methodologies.

Weaknesses

1. The reliance on Markov chains for predictive modeling is a notable concern. Given the complexity of human psychology and the statistical nuances in password generation, Markov chains might not fully encapsulate the breadth of real-world password selection behaviors (specifically, in a semantic context). This limitation should be carefully discussed.

2. This paper claims that the proposed methods are "memory-friendly" and boast "high speed," yet it falls short of providing an exhaustive evaluation of computational resource consumption. A detailed memory utilization and processing time analysis is essential to evaluate the model's practicality and scalability.

3. The absence of source code significantly impedes the study's reproducibility.